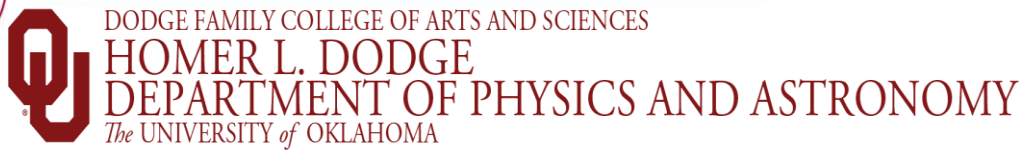
Research Computing and Storage Strategy at the University of Oklahoma

Henry Neeman, University of Oklahoma



I Love Questions!

- I love questions, so please ask questions whenever you think of them!
- I'm more fun to listen to, and I have more fun talking, if we have a conversation, than if I just lecture at you.

So **DON'T BE SHY** – interrupt me all you want!

Also, I have way too many slides, so I'm not expecting to get through everything anyway ….

Forward Looking Disclaimer

- This talk has slides with roadmap items.
- Anything that hasn't been tested doesn't work, by definition.
- These are goals, not guarantees.
- Timelines are extremely fluid.

Outline

- General Information
- Resources
- Supercomputer
- Supercomputer Storage
- Improving Supercomputer **Storage** Reliability Improves **Supercomputer** Reliability Tremendously
- OU Research Cloud (OURcloud)
- OU Research Disk (OURdisk)
- OU & Regional Research Storage (OURRstore)
- OneOklahoma Friction Free Network (OFFN)

General Information

OSCER

- OU Supercomputing Center for Education & Research
- Division of OU Information Technology
  - ~2/3 of institutional research computing organizations report to the CIO, ~1/4 to the VP/VC for Research, ~1/12 other
- Collaboration among dozens of academic and non-academic units across OU's 3 campuses (Norman, Health Sciences, Tulsa)
- Serve not only researchers and educators at OU but also:
  - Their collaborators worldwide
  - Researchers and educators across Oklahoma

OSCER Charging Approach

- If it's **shared** and **temporary**, researchers **don't pay** for it.
  - Instead, it's sponsored by OU's CIO (usually), or funded externally.
- If it's **dedicated** and **persistent**, researchers **buy** it.
  - OU IT maintains it at **no additional charge**, sponsored by OU's CIO.

Research grants typically are:

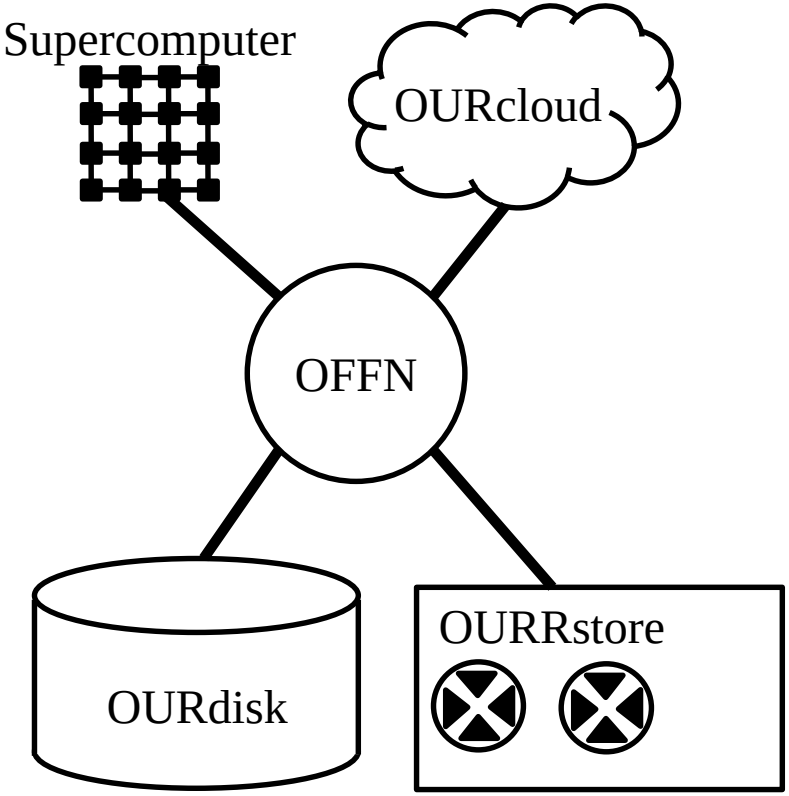
- a **good fit** for hardware/software **purchases**;
- a **poor fit** for **recurring** service charges (in most cases).

So, OSCER favors **one-time, up-front purchases**,
for example from grants.

Resources

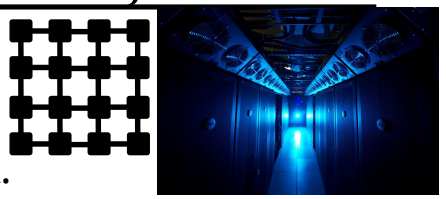
OSCER Resources

Constantly Upgrading Our 5 Major Systems!

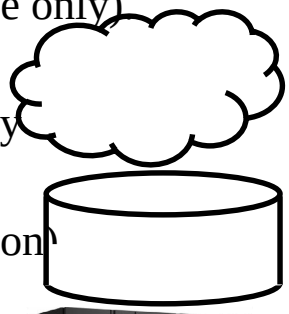


OU Research Computing Summary

OSCER = OU Supercomputing Center for Education & Research, an OU IT team



- **Supercomputer Refresh**
  - Already ~2.1 quadrillion calculations per second (~2.1 PFLOPs) peak.
    - World's fastest: Frontier @ Oak Ridge ~1.7 EFLOPs peak (~1.7 quintillion calc per sec)
  - Lots of CPU, some GPU for Machine Learning, 1000+ TB disk (short term use only)
- **OU Research Cloud (OURcloud)**
  - ~2 TB RAM, 336 virtual CPU cores (@ 3:1 oversubscription), can grow plenty
- **OU Research Disk (OURdisk)**
  - ~8 PB usable, 14+ GB/sec @ OU Norman, ~3.8 PB @ OUHSC (deploying soon)
    - Each of OU Norman and OUHSC can straightforwardly grow to ~22 PB usable.
- **OU & Regional Research Store (OURRstore) Tape Archive**
  - ~11,000 tape cartridge slots now, ~18,000 soon (able to hold 200+ PB).
  - Most HW & SW funded by a National Science Foundation grant.



https://www.pmddatasolutions.com/admin/resources/products/ts4500expanded650x433px-w534h356.gif

- **OneOklahoma Friction Free Network** (OFFN):
  Local and statewide "Science DMZ," research only, 100 Gbps (400 Gbps proposal in preparation).

Supercomputer

Supercomputer Specs

Peak speed: ~2.1 PFLOPs* [@ max turbo/boost] – not including coming deployments

*PFLOPs: quadrillion calculations per second

http://www.oscer.ou.edu/supercomputer

CPU: 29,700 cores + 512 **coming**
- Intel Xeon: Sapphire Rapids, Ice Lake, Cascade Lake, Skylake, Broadwell, Haswell
- AMD EPYC: Rome, Milan, Genoa

GPUs (currently all NVIDIA): 83 + 34 **coming**
- 12 H100, 4 L40S, 16 RTX 6000 Ada, 49 A100, 2 V100; **coming** 24 L40S, 10 H100

~97 TB RAM
~0.9 PB global public disk (+ ~1.2 PB **coming**)
~8 PB OURdisk + ~3 PB condo standalone disk
NVIDIA/Mellanox Infiniband ~1 microsec latency
- (FDR10 3:1 oversubscribed, 13.33 Gbps,
- HDR100 4:1 oversubscribed, 25 Gbps)

Dell N-series Gigabit, S-series 25G/100G Ethernet
NVIDIA/Mellanox "Skyway" IB-to-Eth gateway × 2
Enterprise Linux, currently upgrading to 9.4
Around half of the nodes are "condominium" (owned by individual research teams).

schooner.oscer.ou.edu

Photo: Jawanza Bassue

Node Flavors

- Compute
  - Regular (CPU-focused)
  - GPU
  - Large RAM
- Support
- Storage

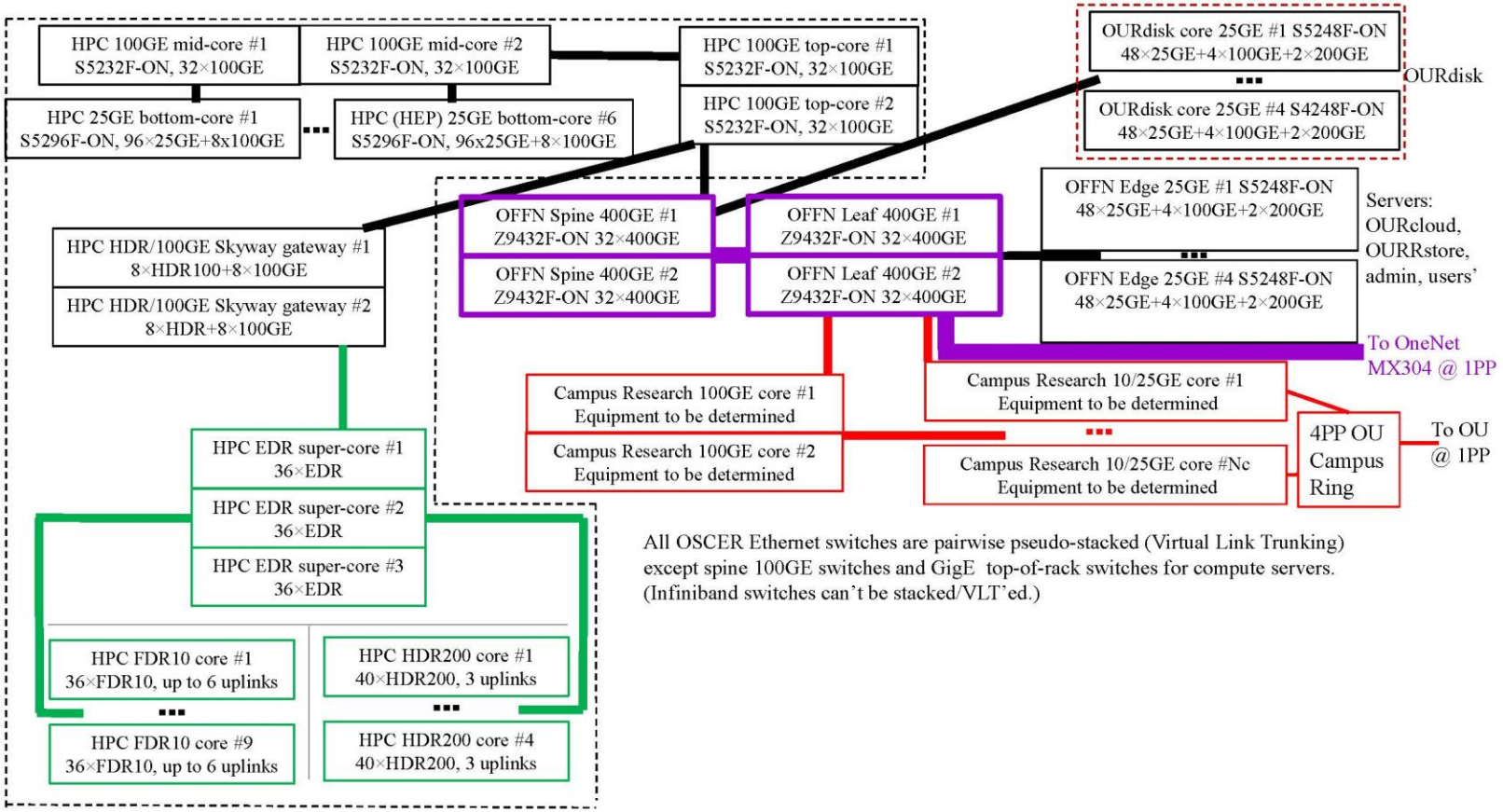
Supercomputer Condominium Nodes

- **Buy**: OU users can buy "condominium" nodes any time!
  - The researcher buys the node and a few network cables.
  - OU's COI sponsors space, power, cooling, network (including internal networks) and labor.
  - The condominium node remains in production for the lifetime of the current supercomputer and its immediate successor.
- **Researchers' Pricing**: available on request (OSCER requests the quote from the vendor, based on our standard configuration).
- **Flavors**: Compute nodes, GPU nodes, large RAM nodes
  - Can change CPU model, RAM size
- **Storage**: See OURdisk (slides a little later).

Supercomputer Networks

- Internal Dedicated
  - Infiniband (IB)
    - Older Nodes: FDR10: 40 Gbps, 3:1 oversubscribed
      => 13.3 Gbps when fully saturated, 9 core switches)
    - Newer Nodes: HDR100: 100 Gbps, 4:1 oversubscribed
      => 25 Gbps when fully saturated, 4 core switches)
    - Bridge between FDR10 and HDR100: EDR (3 super-core switches)
  - Ethernet
  - IB-to-Ethernet
- World-Facing
  - OneOklahoma Friction Free Network (OFFN)
    - Oklahoma's statewide Science DMZ

Supercomputer Networks



All OSCER Ethernet switches are pairwise pseudo-stacked (Virtual Link Trunking) except spine 100GE switches and GigE top-of-rack switches for compute servers. (Infiniband switches can't be stacked/VLT'ed.)

Supercomputer Infiniband

- NVIDIA/Mellanox FDR10: 40 Gbps
  - On most of our older nodes
  - 3:1 oversubscribed => 13.3 Gbps when fully saturated
  - 9 FDR10 core switches
  - Replacing all of this would have been low-to-mid 6 figures!
- NVIDIA/Mellanox HDR100: 100 Gbps
  - On most of our newer nodes
  - 4:1 oversubscribed => 25 Gbps when fully saturated
  - 4 HDR[200] core switches
- NVIDIA/Mellanox EDR: 100 Gbps
  - Bridge between FDR10 and HDR100
  - 3 EDR super-core switches

Supercomputer Ethernet #1

- Each node: GigE for management, and for nodes that lack Infiniband, also disk I/O (common) and MPI (rare).
- Some nodes: 25GE for disk I/O.
  - This includes many HEP nodes.
- Some nodes: 25GE as failover in case IB fails.
  - GPU nodes (a few dozen)
  - Large RAM nodes (a handful)
  - Support nodes
  - Diskfull nodes
- A few very special nodes: 100GE
  - Data mover nodes and a few others

Supercomputer Ethernet #2

- Top-of-rack GigE 48-port switches, mostly for management.
- 4 × 25GE core switches
  - For nodes with 25GE cards and for top-of-rack GigE switches
- 2 × 100GE super-core switches (soon 4)
  - For uplinking 25GE core switches
  - For connecting the supercomputer to the world
    - Via the OneOklahoma Friction Free Network (OFFN)

IB-to-Ethernet Gateways

- 2 × NVIDIA/Mellanox “Skyway” Gateway Appliances
  - Each Skyway has 8 × HDR + 8 × 100GE.
  - The pair of Skyways is in High Availability mode.
- Why have an IB-to-Ethernet gateway?
  - Some nodes have IB, some only have Ethernet.
  - Some storage has IB, some only has Ethernet.
  - Every node needs to be able to talk to all storage.
    - This is especially relevant for OURdisk (coming up), which uses physical Ethernet only, no IB.

Supercomputer Software

- OS: mostly Enterprise Linux 9, some EL8, transitioning off of EL7 now.
  - Mostly Rocky.
- Scheduler: Slurm
- Compilers: Intel OneAPI, GNU, NVIDIA (formerly PGI)
- Parallel debugger: TotalView
- Interactive Use: Open OnDemand
- Management: Salt, Paramiko
- Ethernet Switches: Dell Enterprise OS10, OpenNMS
- Monitoring: Node Health Check, Nagios
- Measurement: Telegraf, Prometheus
- Instrumentation: Grafana, XDMoD

Supercomputer Storage

Storage: /home

- **/home**: For software packages, scripts, small input files.
  - **EVERY USER** gets a /home directory.
  - **NO CHARGE**: Sponsored by the CIO.
  - **PERSISTENT**: Files remain there unless the user deletes them.
  - **SMALL**: Typically a few tens of GB per user.
  - **BACKED UP**: Nightly incremental, occasional full dump.
  - **SLOW**: Server with 8 hard (spinning) drives.
  - 2 /home subsystems, each with half of the supercomputer's users.
    - If one of the /home subsystems fails, the other one is still up, so the supercomputer is still up: jobs keep running, but from only half the users, instead of all the users.

Storage: /scratch

- **/scratch**: For bulk datasets.
  - **EVERY USER** gets a /scratch directory.
  - **NO CHARGE**: Sponsored by the CIO.
  - **TEMPORARY**: Files remain there for 2 weeks.
  - **LARGE**: No quota limit; the only limit is the physical capacity of the storage device, which is shared among hundreds of users (but usually only a few tens of users at a time).
  - **NEVER EVER BACKED UP**
  - **FAST or MEDIUM SPEED**: 2 flavors:
    - **Parallel /scratch**: CephFS, 7 diskfull servers @ 20 × 20 TB HDDs, erasure coding (4 data chunks + 2 redundancy chunks, server-level).
    - **IOPS /scratch**: servers full of hard disk drives (up to 16 HDDs per server, NFS-on-ZFS on top of physical RAID6 + hot spare).
      - Soon to be 4 NFS-on-ZFS /scratch subsystems, covering ~80% of users.

Why 2 Flavors of /scratch?

- **Sequential**: Some users do lots of sequential I/O, not much IOPS.
  - Example: weather forecasting:
    - Each of 1000 CPU cores does 5 minutes of heavy number crunching.
    - Then every core writes 100 MB to disk, maybe even to the same file, in parallel.
    - Then they all do it again, over and over.
  - These users need a filesystem that's optimized for sequential I/O.
    - This is what high performance parallel filesystems are designed for.
- **IOPS**: Some users do random I/O: zillions of tiny writes/reads.
  - These users need a filesystem that's optimized for IOPS.
- **Both**: Some users do both sequential I/O and IOPS.
  - They need a filesystem that's big, and has both a decent IOPS rate and adequate sequential bandwidth (NFS-on-ZFS).

Metadata/Reread Cache on SSD

- ZFS allows metadata to be on SSD, even if the data is on HDD.
- ZFS also supports using SSD for reread cache
  (but not for write cache, unfortunately).
- We're about to test this on one of our IOPS /scratch subsystems.
- Assuming it's successful there, we'll also put it on
  both of our /home subsystems and all of our IOPS /scratch
  subsystems.
- We expect metadata on SSD to improve read IOPS performance.
  - Not so much for write IOPS performance, because
    the application doing the I/O considers a write to have completed
    once the write gets to diskfull server RAM.

Storage: Burst Buffer

- Burst Buffer: Server with 16 × NVMe SSDs + 8 × HDR100 ports
  - Can be reserved by capacity for each batch job.
  - Ideal for large numbers of small reads: auto-stage-in your input files to burst buffer, do your small reads there, auto-stage-out your output.
  - Partially tested, not yet in production.

Storage: OURdisk & OURRstore

Slides about the following are coming up shortly, but these **aren't properly part of our supercomputer**.

- **OURdisk** is persistent, dedicated disk storage that users can buy into.
  - OURdisk is mounted on our supercomputer, but OURdisk **isn't part of** our supercomputer (it can also be mounted on OURcloud, on servers, and on PCs).
- **OURRstore** is our tape archive.
  - OURRstore is mounted on a few nodes of our supercomputer, but OURRstore **isn't part of** our supercomputer (it's also accessible from outside our supercomputer).

Improving Supercomputer Storage Reliability Improves Supercomputer Reliability Tremendously

Storage Goals

- "Big, fast and cheap – pick any two!"
- We actually want:
  - big;
  - fast;
  - cheap;
  - reliable.
- How can we accomplish all four of these at the same time?

Why Do Bank Robbers Rob Banks?

- In 1923, bank robber Paul Perrit said he specialized in robbing banks "because that's where the money is."
  - **NOTE**: This quote is often attributed to bank robber Willie Sutton, but that attribution started showing up in 1951.
    - https://quoteinvestigator.com/2013/02/10/where-money-is/

- Moral: Figure out where the bulk of the issue is, and focus on that.

Supercomputer Uptime: Biggest Factor

- The **greatest threat** to supercomputer uptime is **storage failures!**
- **79%** of ACCESS and XSEDE cluster supercomputer failures have been **storage failures!**
  - May 25 2012 – Nov 2 2024: https://support.access-ci.org/outages
  - 13 different cluster supercomputers
    - Anvil, Bridges, Bridges2, Comet, Delta, Expanse, Gordon, KyRIC, Lonestar, Maverick, Stampede, Stampede2, Steele
    - We went through the entire dataset, which goes back to 2008 (cluster supercomputer unscheduled outages start May 25 2012).
    - We excluded cloud resources like Jetstream and OSG.
  - Lots of variation on the number of incidents year by year, but no trend.
- So, improving supercomputer **storage** uptime improves **supercomputer** uptime **tremendously**!

OU's Approach to Improving Supercomputer Storage Reliability

Multiple Storage Subsystems per Type

- We have multiple storage subsystems of each type:
  - 2 /home subsystems;
  - 5 /scratch subsystems (1 CephFS, 4 NFS-on-ZFS).
    - As described below, 1 in production now, 3 more soon.
- If one of the /home or /scratch subsystems crashes, then its subset of users are becalmed – but everyone else is fine!
  - Downtime for some, but the supercomputer stays full, because there are plenty of jobs pending in the batch queues to replace the jobs that crashed because their storage crashed.
    - If Kim's running job crashes because Kim's /home or /scratch crashes, Lee's pending job starts, so the supercomputer stays full!
- In other words, a failure of one storage subsystem isn't a failure of the supercomputer as a whole, because the supercomputer stays full.

DRBD

Distributed Replicated Block Device (DRBD) https://en.wikipedia.org/wiki/DRBD

- Added to the Linux kernel in late 2008.
- We'll use it for NSF storage (/home and NFS-on-ZFS /scratch).
- Every such server will actually be a pair of (nearly) identical servers, one a primary and the other a secondary (like RAID1).
- Each disk write goes to the primary and then to the secondary, in "write-through" mode: a write is complete only after it commits to the secondary diskfull server.
  - Write IOPS are significantly slower than without DRBD, **BUT**
  - if a diskfull server fails, the other one in the DRBD pair continues, and has all the files: no data loss, and the user doesn't even notice!
- DRBD is in production on one of OSCER's NFS-on-ZFS /scratch subsystems (it's working great), and soon we'll have DRBD on all NFS-on-ZFS /home and /scratch subsystems.

OU Research Cloud (OURcloud)

OU Research Cloud (OURcloud)

- **Purpose**: Interactive, web services, databases (e.g., SQL), Windows OS, etc.
  - **http://www.oscer.ou.edu/ourcloud**
- **Researcher's Price**: $347.19 per portion (minimum buy-in)
  - Portion: 16 GB RAM + 2 virtual CPU cores (=> 2/3 physical CPU core @ 3:1 oversubscribed), in production for **7 years**
  - Ram is 7/8 undersubscribed (for OS + I/O buffering).
  - Least expensive research cloud offering in OU IT history!
  - Can also buy condominium servers (OU CS has done this).
- **Size**: Currently 336 portions (~3 TB RAM, 672 virtual CPU cores) – will grow as needed (+0.75 TB purchased, arriving soon).
- **OS Options**: Linux (multiple versions), Windows Datacenter (most recent version).

OU Research Disk (OURdisk)

OU Research Disk (OURdisk) #1

- **Purpose**: Persistent, dedicated disk space on supercomputer, OURcloud, other servers across OU, PCs – **http://www.oscer.ou.edu/ourdisk**
- **Researcher's Price**: $860.03 (minimum buy-in) per 9.3 usable TB portion, in production for **7 years** => ~$93 per usable TB => ~$0.001/GB/month
  - Least expensive research disk offering in OU IT history!
- **Speed**: 14+ GB/sec aggregate
  - Fastest research spinning disk offering in OU IT history!
  - Individual sequential write: ~1 GB/sec
- **Size**: Currently ~8 PB @ OU Norman, soon ~3.8 PB @ OUHSC
  - **Already at OU Norman**: ~7.75 PB bought, ~5.75 PB consumed
    - Can add enough servers to reach ~22 PB at each of OU Norman & OUHSC
    - 77 OURdisk research teams since Nov 2020 - Oct 2024
      (vs 12 groups buying condominium standalone storage 2012-2020)
- **Where Available**
  - Supercomputer, OURcloud, external servers, PCs
  - Can be mounted on OU IT systems and non-IT systems on any OU campus.
- **Software**: CephFS

OU Research Disk (OURdisk) #2

- Each of OU Norman and OUHSC will have 2 independent systems.
- A user can buy portions on multiple standalone OURdisk systems.
  - We expect this option to be unpopular because of cost.
  - But, it enables proper backups or mirroring (can't use OURRstore).
- **Capacity**: ~8 PB usable OU Norman (soon ~3.8 PB @ OU Health Sciences Ctr): 33 × diskfull server @ 24 × 16/18/20 TB spinning drives + 2 × SSD for metadata; 5 × metadata servers
  - Each campus's storage capacity will grow with demand on that campus.
- **Resiliency**: 8 + 3 server-level "erasure coding" (better than RAID6), so up to 3 simultaneous failed servers or drives would be invisible to users.
  - We wrote a disk drive failure Monte Carlo simulator that showed many double failures, very few triple failures (0.1% chance in 5 years), **ZERO** quadruple failures.
- **Network**: 2 × 25GE uplinks per server, both diskfull and metadata (plus GigE connections per server for management) – 24 × HDD => ~14.4 Gbps
- **Science DMZ Research-only Network**: OU Norman 25GE switches uplink to 100GE OneOklahoma Friction Free Network (OFFN) switches; similar at OUHSC soon.

Why Ceph for OURdisk?

Ceph is the only Software Defined Storage technology that has all of:

- **FREE** and open source.
- Parallel filesystem optimized for sequential I/O.
- Can be big: up to ~1000 spinning drives per Ceph system, meaning up to ~14 PB per Ceph system (Norman OURdisk #1: 792).
  - If all drives are 24 TB, because 28-32 TB are all Shingled (slow).
- No specialized components: works on pretty much any hardware.
- Relatively straightforward to manage
  (though the learning curve was **painful**).
- Any server, whether diskfull or support, can be replaced with little or no downtime, so each server can be used for its natural lifetime, then decommissioned, but the Ceph resource as a whole **CAN LAST INDEFINITELY**.

FS-Cache #1

- FileSystem Cache (FS-Cache) is a longstanding part of the Linux kernel:
  - https://docs.redhat.com/en/documentation/red_hat_enterprise_linux/6/html/storage_administration_guide/ch-fscache#ch-fscache
- FS-Cache allows I/O transactions to be cached to one filesystem on their way to/from another.
- FS-Cache works well with the cache on NFS-on-XFS (or EXT4) on physical RAID.
- FS-cache works well with DRBD across pairs of servers.
- We use read-intensive SATA SSDs: cheap, middling bandwidth (~0.5 GB/sec per SSD) but reasonable IOPS rates (30-43K write, 79-92K read per SSD) – we can aggregate many such SSDs affordably, to get decent total SSD footprint (currently ~12 TB).

FS-Cache #2

- FS-cache is **write-back for data** (data lands on FS-cache, then drains gradually to the far target), but **write-through for metadata** (as soon as you write, the metadata goes immediately to the far target).
- FS-cache is valuable for write IOPS, but not for read IOPS (for read IOPS, it actually makes transactions a bit slower).
- FS-cache isn't worthwhile for scratch, because we'd burn through lots of SSDs: growing towards ~25 TB/day of scratch files, meaning we'd consume all the SSDs' write endurance (drive writes per day) and have to buy more.

FS-Cache: Supercomputer-to-OURdisk

To speed up writes to OURdisk from the supercomputer:

- DRBD pairs of servers full of 960 GB SATA SSDs:
  - 1 pair at 16 × 960 GB SATA SSDs each (can grow to 32) – testing this now;
  - 3 pairs at 8 × 960 GB SATA SSDs each (coming).
- Each supercomputer-to-OURdisk FS-cache DRBD pair has dual Infiniband HDR100 (IP-over-IB), plus dual 25GE for failover.
- Automount allows auto-failover from one FS-cache DRBD pair to another, so it's okay if a DRBD pair fails.
- Doesn't burn through lots of SSDs, because OURdisk users mostly write once, rarely delete, so relatively few PB.
  - Aggregate FS-cache write endurance will be 50+ PB.

FS-Cache: Far-Building-to-OURdisk

Remote in-the-building FS-cache: Some buildings remote from OSCER's primary data center ("4PP") will have a local FS-cache in the building.

- Nielsen Hall (Physics & Astronomy), National Weather Center, and 2 network aggregation points
- Disk I/O will pass through FS-cache, then drain off to OURdisk in 4PP afterwards.
- For network aggregation points, firewalls are on the far side.
- Fixed cost per such building, not much proportional to traffic.
- Limiting factor: rack space, power, cooling and network ports are in short supply in most campus buildings.
- Already purchased, waiting to be deployed.

OU & Regional Research Store (OURRstore)

OURRstore Tape Archive #1

- **OU Regional & Research Store: Giant robotic tape archive**
  - Business Model: NSF MRI bought HW/SW, researchers buy tapes, OU's CIO covers space/power/cooling/network/labor/maintenance.
    - Currently ~$62 per LTO-8 tape cartridge (~10.2 TB usable)
      => ~$12 per usable TB for dual copies
  - Tape Cartridge Slots: Initially ~11,000, expanding to ~18,000 soon (~11K @ Norman, ~7K @ OUHSC)
  - Tape Drives: Initially, ~1.8 GB/sec in aggregate – almost double PetaStore!
    - 6 × LTO-8 @ 360 MB/sec/drive
  - Disk: ~570 TB usable disk front end "landing pad."
  - Software: IBM Spectrum Archive for tape,
    IBM Storage Scale/Spectrum Scale/GPFS for disk.
  - Resiliency: Secondary copies exported from OURRstore, shelved or shipped.

https://www.pmddatasolutions.com/admin/resources/products/ts4500expanded650x433px-w534h356.gif

OURRstore Tape Archive #2

Purchased, waiting to be delivered/deployed:

- **NEW!** 2nd tape library @ OU Health Sciences Center in OKC
  - 1 × L55 control frame, 5 × S55 cartridge-only expansion frame
  - Brings OURRstore's total capacity to ~18,000 tape cartridge slots.
- **NEW!** 6 × LTO-9 tape drives: new total bandwidth ~4.5+ GB/sec
  - Planning to buy 6 × LTO-10 tape drives in ~2027.
- **NEW!** Disk front-end landing pad for files coming on and off
  - IBM FlashSystem 5300: 12 × 9.6 TB NVMe SSD, 72 × 12 TB NLSAS
    - And can incorporate our extant old FlashSystem 5030 too!
- **NEW!** Servers (x86)
  - 4 servers for tape control (2 per tape library)
  - 4 servers for disk control
- **NEW!** Software licenses (Spectrum Archive, GPFS)

https://www.pmddatasolutions.com/admin/resources/products/ts4500expanded650x433px-w534h356.gif

OURRstore Roadmap

The LTO roadmap that goes to LTO-14.

https://www.businesswire.com/news/home/20240901005862/en/LTO-Program-Announces-Extension-to-the-LTO-Tape-Technology-Roadmap-to-Generation-14

- Best guess timing estimates
  - LTO-10: drives c. 2024, cartridges breakeven $/TB c. 2028
  - LTO-11: drives c. 2027, cartridges breakeven $/TB c. 2031
  - LTO-12: drives c. 2030, cartridges breakeven $/TB c. 2034
  - LTO-13: drives c. 2033, cartridges breakeven $/TB c. 2037
  - LTO-14: drives c. 2036, cartridges breakeven $/TB c. 2040
    - I'll be retired!
- OURRstore will be in production until at least 2030.
- The new LTO roadmap gives us a post-OURRstore plan that can take us to ~2042 (almost 2 more decades), when 1 PB ~ 1 lb.

Why Tape?

- Tape sucks. We all know it. Everyone hates tape.
- The problem is, the alternative to tape isn't disk; the alternative to tape is deleting all your data.
- The reason we like tape is, tape is dirt cheap: LTO-9 tape cartridges are less than half the price per TB of USB drives from a big box store (let alone the price of enterprise-class storage systems).
- So tape is awesome, even though it sucks.

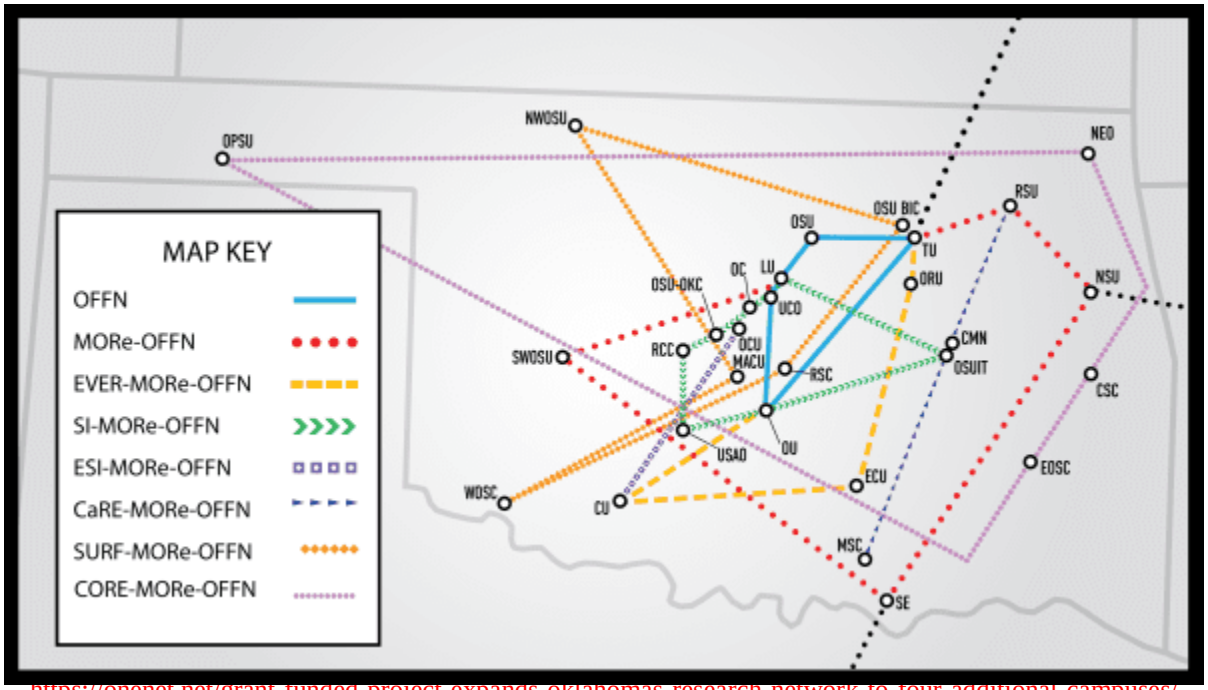
OneOklahoma Friction Free Network (OFFN)

OneOklahoma Friction Free Network

- **Researcher's Price**: **ZERO** (sponsored by NSF, OU's CIO)
  - Originally funded under NSF Campus CI grant in 2013.
  - 400GE proposal submitted Oct 15 2024.
- Science DMZ: High speed network for **open** research only.
  - Friction Free: Bypasses firewall appliances because the data is open.
    - Firewalling without firewall appliances allows much higher speed, because firewall appliances interpret research data flows as attacks.
- Funded statewide by 9 NSF Campus Cyberinfrastructure grants.
  - 30 institutions
  - First 2 grants led by OU
- OU's open research connection to:
  - Other Oklahoma research institutions
    - 30 institutions: PhD-granting, masters-granting, bachelors-granting, community colleges, Minority Serving, Tribal
  - Research institutions across the US

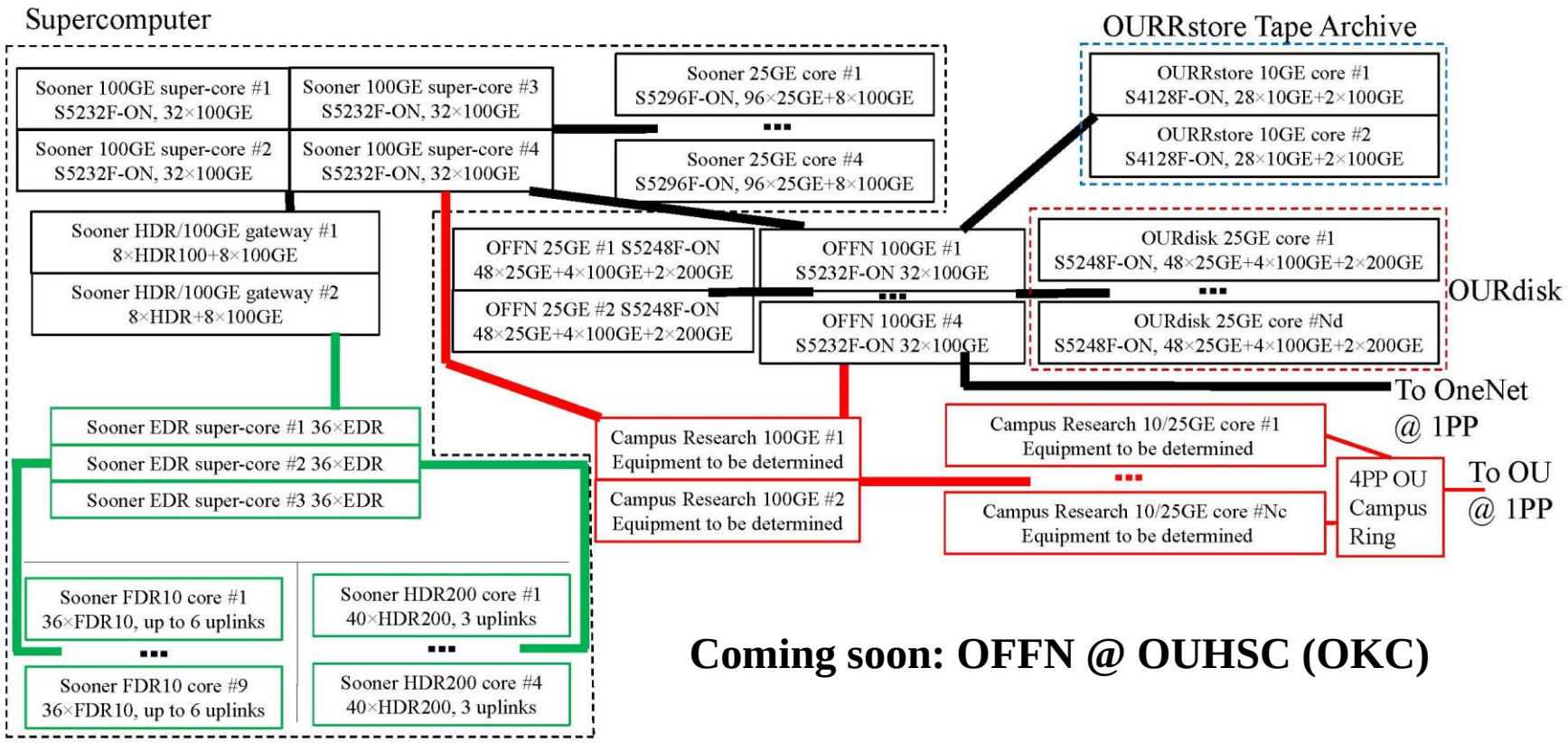
OFFN Across Oklahoma

9 OFFN NSF CC* grants
(OU led the first 2)
30 institutions



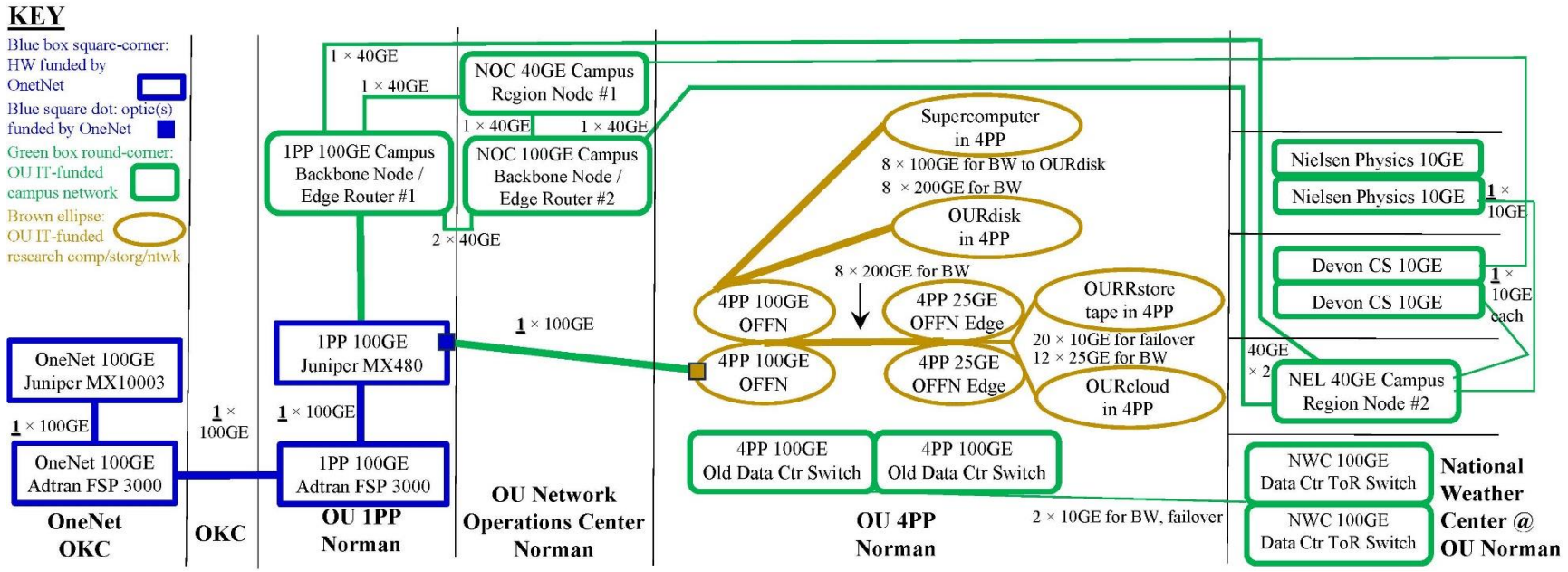
https://onenet.net/grant-funded-project-expands-oklahomas-research-network-to-four-additional-campuses/

OFFN @ OU Norman #1



Coming soon: OFFN @ OUHSC (OKC)

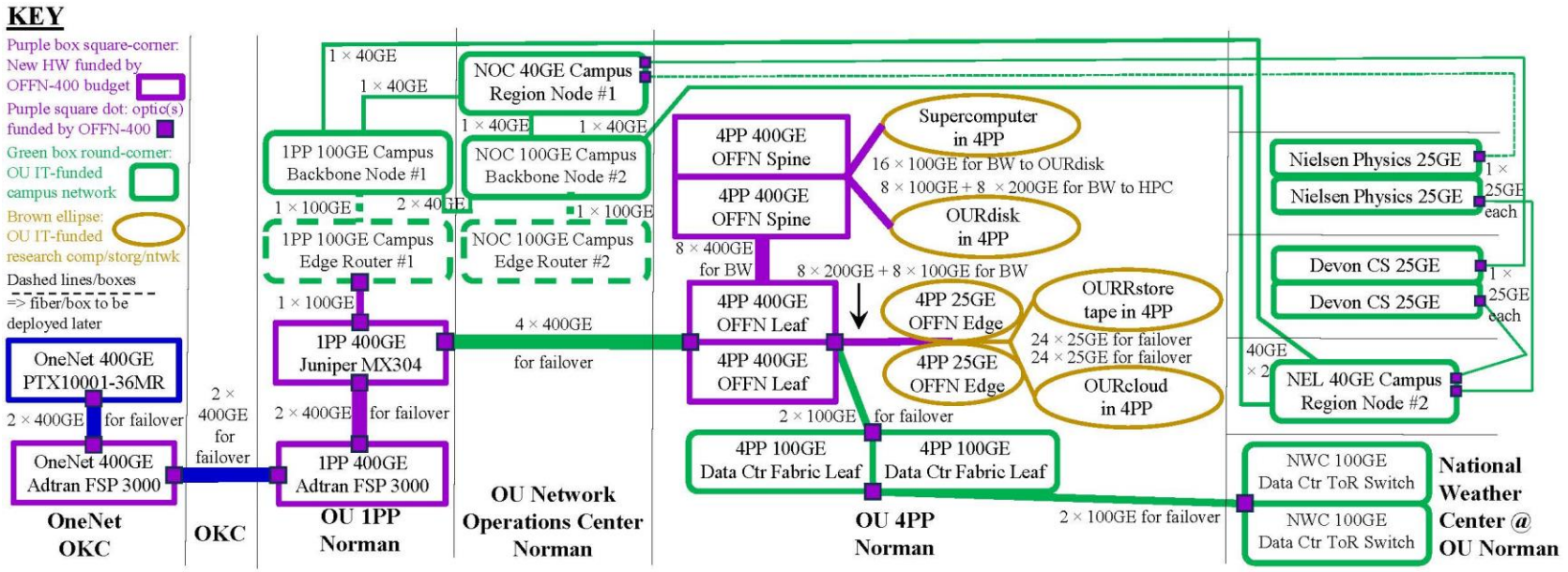
OFFN @ OU Norman #2



Proposed: OFFN 400GE @ OU Norman #1

NSF CC* proposal, submitted Tue Oct 15

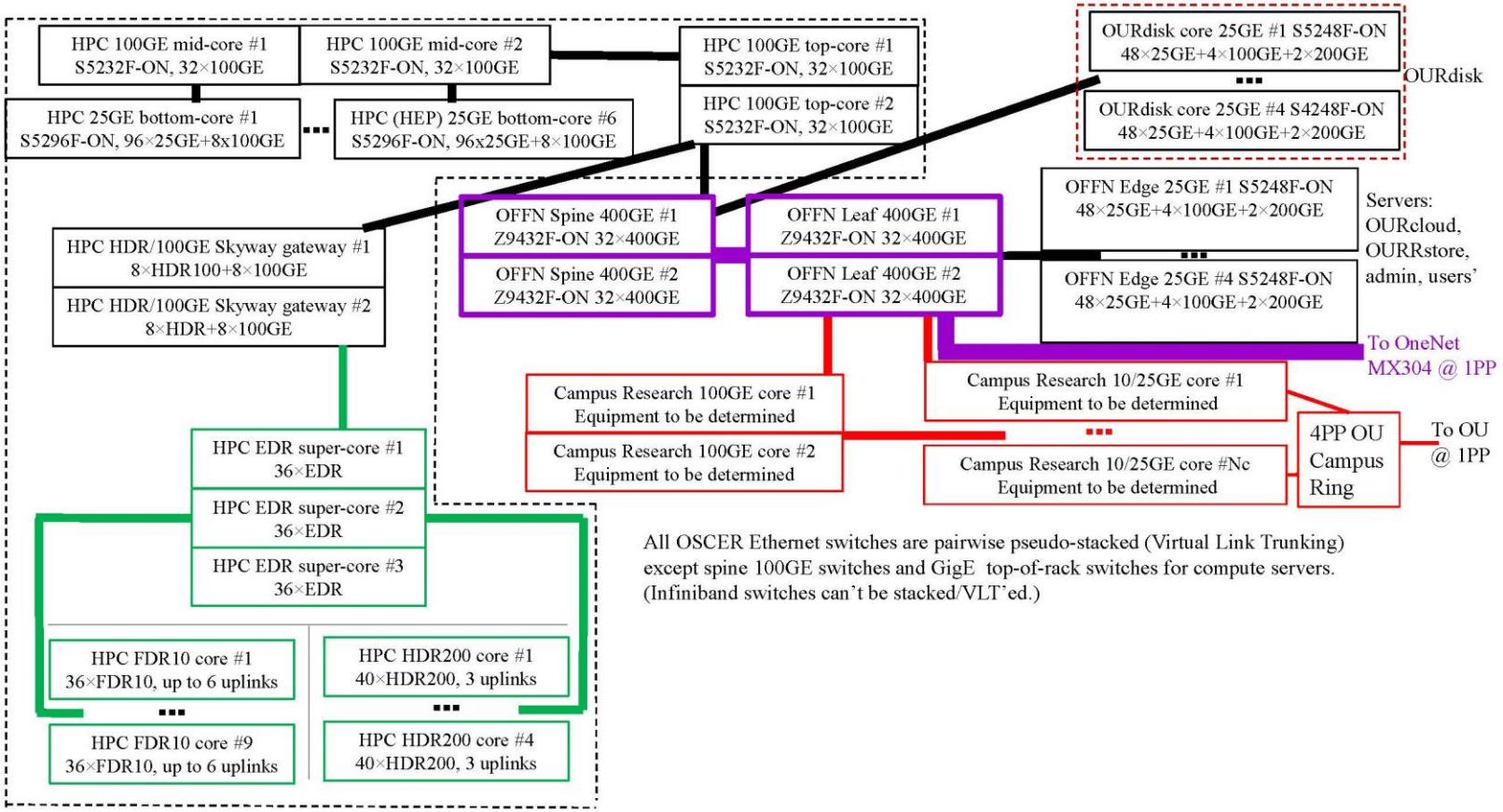
31 research teams have signed on, dozens more have expressed interest



Coming: OFFN 400GE @ OU Norman #2

NSF CC* proposal, submitted Tue Oct 15

31 research teams have signed on, dozens more have expressed interest



All OSCER Ethernet switches are pairwise pseudo-stacked (Virtual Link Trunking) except spine 100GE switches and GigE top-of-rack switches for compute servers. (Infiniband switches can't be stacked/VLT'ed.)

Thanks for your attention!

Questions?